ABSTRACT OF THE DISCLOSURE

A technique for determining when documents stored in digital format in a data processing system are similar. A method compares a sparse representation of two or more documents by breaking the documents into "chunks" of data of predefined sizes. Selected subsets of the chunks are determined as being representative of data in the documents and coefficients are developed to represent such chunks. Coefficients are then combined into coefficient clusters containing coefficients that are similar according to a predetermined similarity metric. The degree of similarity between documents is then evaluated by counting clusters into which chunks of similar documents fall.